Charset e Internet

Una introduzione sulle problematiche delle codifiche di carattere in Internet

(e non solo)

Maurizio Manetti

2012

Problemi tipici...

- Vedo caratteri “strani”
- Scrivo è e leggo Ã¨, perchÃ© ?
- Scrivo è e leggo , perch?
- Vedo dei quadratini al posto delle lettere
- Vedo i caratteri accentati con font diversi
- Ma non era una virgoletta singola? ‘ ` ' ’ ‘ ′ ´
- Ma non era una virgoletta doppia? “ ” " „ ″
- Ma non era un trattino? - – — — –
- Salvo il file caffè.txt e in FTP vedo caff?.txt
- Non vedo le lettere accentate nella bash
- Non riesco a estrarre sottostringhe di lunghezza voluta
- Mi arrivano le email con č al posto di è
- Ho aperto un file di testo e comincia con ï»¿
- Il cliente vuole il sito in giapponese... funzionerà con la piattaforma attuale? Che succede alle URL?

Joel Spolsky

If you are a programmer working in 2003 and you don't know the basics of characters, character sets, encodings, and Unicode, and I *catch* you, I'm going to punish you by making you peel onions for 6 months in a submarine. I swear I will.

http://www.joelonsoftware.com/articles/Unicode.html

The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!)

Unicode

The Unicode Consortium Members

Full Members

Adobe
Google
IBM
Microsoft
ORACLE
REARDEN COMMERCE
SAP
YAHOO!

Institutional Members

Government of India
ভাষা প্রযুক্তি গবেষণা পরিষদ
Society for Natural Language Technology Research
Berkeley

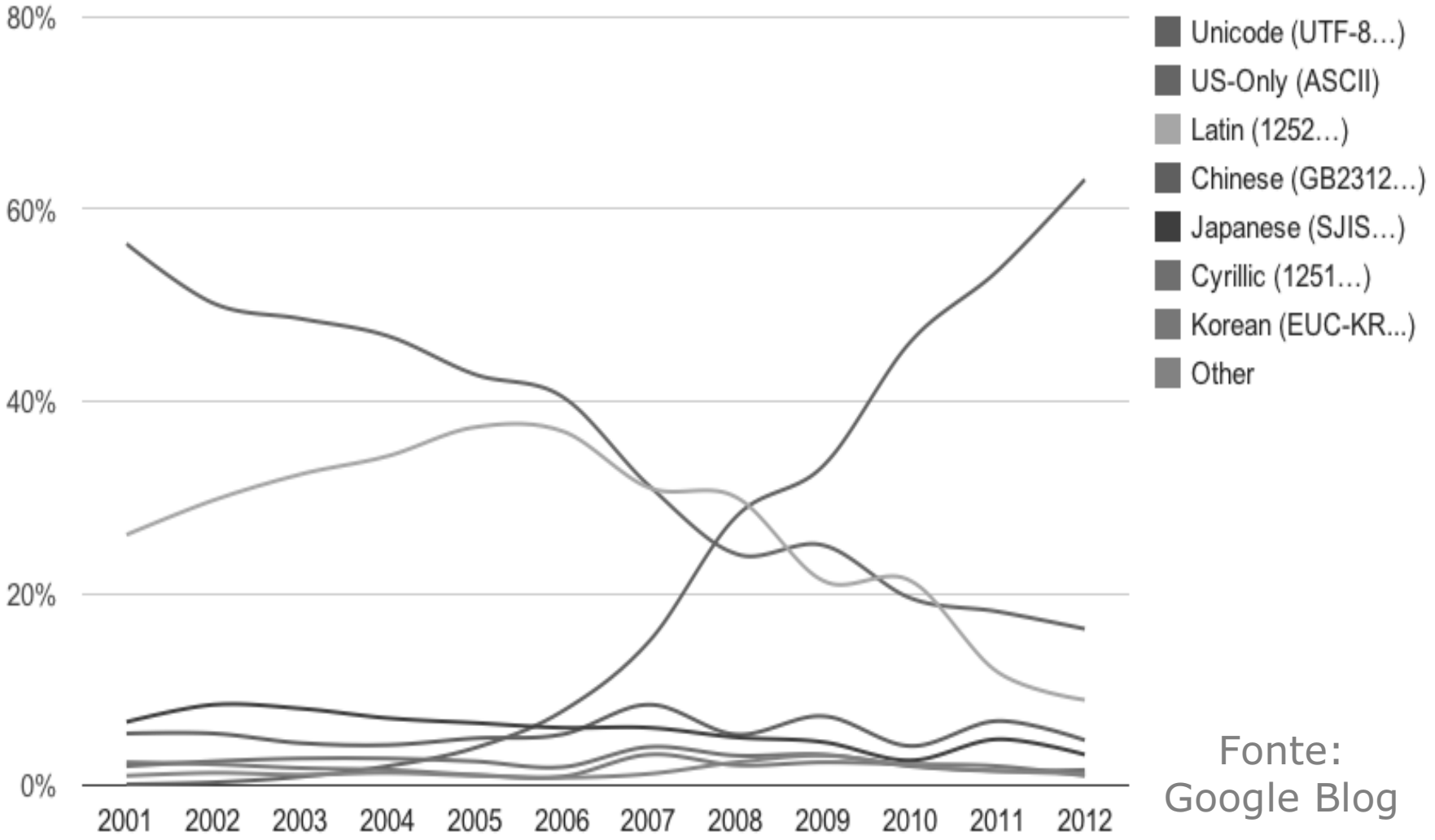
Supporting Members

Monotype Imaging

Associate Members

BASIS TECHNOLOGY
BEIJING ZHONGYI ELECTRONICS CO.
THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS
compart
core Internet Council of Registrars
DECO type
DENIC
digicert
evercype
HIGHTECH PASSPORT
INNOVATIVE
Linotype
NOKIA
OCLC The world's libraries. Connected.
The perl Foundation
SANMYANMAR
sas
SIL International
Sony Ericsson
Tavultesoft
TERADATA Raising Intelligence
IT Direct DATA QUALITY SOLUTIONS
Utimka
VeriSign
文林 Wénlín

Uso di Unicode nel Web



Fonte:
Google Blog

Questioni di charset

- Filesystem
- Clipboard (copia e incolla)
- File di testo e word processors
- Database
- FTP
- Stringhe nei linguaggi di programmazione
- Email
- Web (browser / server)
- Dappertutto (ovunque vi sia una gestione digitale del testo)

Concetti chiave

- Bit
- Byte
- Ottetto
- Carattere (e grafema)
- Glifo
- Font
- Script (writing system)

Bit

- 2 significati:
  - binary unit (quantità di informazione)
  - binary digit: uno dei due simboli del sistema numerico binario (0,1)
- binary unit: teoria di Shannon (1948)
- binary digit: unità di definizione di uno stato logico (nasce con le schede perforate)

Byte

- 1 byte ≠ 8 bit (storicamente)
- una sequenza di **bit**, il cui numero dipende dall'implementazione fisica della macchina sottostante
- numero di bit utilizzati per codificare un "**singolo carattere di testo**" in un computer
- nibble e word
- Werner Buchholz (1956) fase di progetto IBM Stretch
- respelling di "bite" per evitare confusione con "bit"
- codice Fieldata (U.S. Army & Navy, 1956-1962): 6 bit

Ottetto

- Identifica senza ambiguità un byte composto da 8 bit o, più in generale, un raggruppamento di 8 bit
- Viene utilizzato negli standard e in generale negli RFC per evitare confusione
- Rappresentazioni:
  - Binaria (00000000 - 11111111)
  - Ottale (0 – 377)
  - Decimale (0 - 255)
  - Esadecimale (00 – FF)
  - Per noi 1 byte = 1 ottetto

Carattere

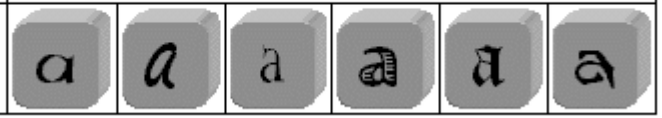
- Concetto sfuggente in informatica e in tipografia
- Forma platonica (non funziona bene: A ≠ a)
- Più facile dire cosa non è un carattere:
  - non è un glifo
  - non è un grafema (unità atomica di uno Script)
  - non è un nome
  - non è un fonema
  - non è una combinazione di bit
- Unità atomica della comunicazione scritta: un simbolo tra le cui varie rappresentazioni c'è un accordo di significato in una determinata comunità di persone
- I problemi nascono quando persone diverse interpretano i simboli in modi diversi

Carattere in Unicode

- Non ha un aspetto determinato: il glifo può variare entro ampi limiti, fintanto che viene riconosciuto
- È essenzialmente bianco e nero, sebbene nel complesso possa essere colorato con una qualsiasi combinazione di due colori (di fatto non è più vero con gli *emoji* in Unicode 6.0, ma solo in combinazione con i *variant selectors*)
- Ha un nome (e una posizione) ufficiale immutabili
- Ha una serie di caratteristiche (categoria, direzionalità, etc..)
- Non ha una pronuncia fissa (tranne alcune eccezioni)
- Può avere utilizzi molto specifici, come i simboli speciali (©) o utilizzi molto vasti per un'ampia varietà di scopi (/)
- Può essere non rappresentabile (control, format, altro..)
- Nell'uso concreto nascono problemi e ambiguità (vedremo nel seguito)

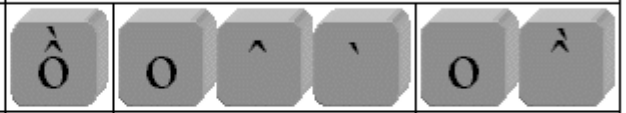
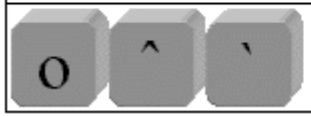
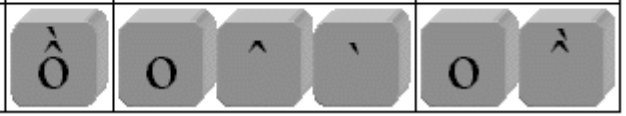
Glifo

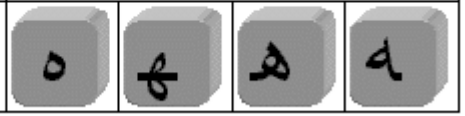
- È un'unità tipografica
- Il suo aspetto dipende da molti fattori (in maniera ovvia ad esempio dal font, ma non solo)
- Si potrebbe dire che è un'istanza concreta di un carattere…
- …ma non è vero! Infatti:
- Può essere composto da più caratteri
- Un carattere può essere composto da più glifi
- Uno stesso carattere può avere diverse rappresentazioni (ovvero glifi diversi per rappresentarlo) in base al contesto e determinate regole nel sistema di scrittura (Script) in cui il carattere esiste (nello stesso font)
- Caratteri diversi possono essere rappresentati dallo stesso glifo (casi particolari, "a" e "alpha" maiuscoli o caratteri di compatibilità)
- Con Unicode le cose si complicano a causa del supporto ai differenti sistemi di scrittura in uso nel mondo

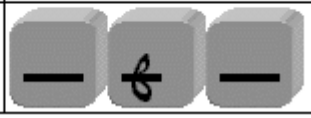
Character	Sample Glyphs
a	

Character Sequence	Sample Glyph
f i	

Character Sequence	Possible Glyph Sequence
f i	

Character Sequence	Possible Glyph Sequences
	
	

Character	Contextual Glyph Shapes
	

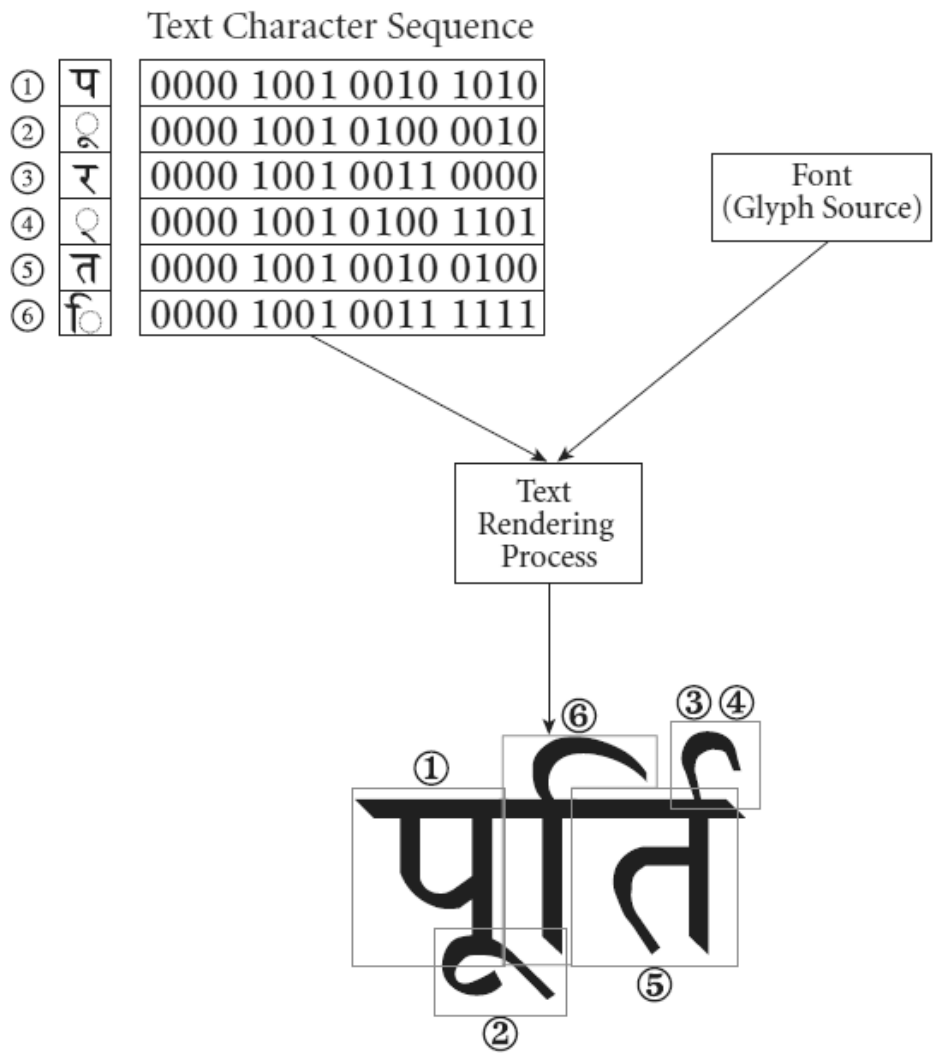
Character	Sequence of glyphs
	

Character	Split Glyphs
	

Font

- Origine dal latino *fundere*, con riferimento ai caratteri mobili per la stampa tipografica
- In generale (in ambito tipografico) ha una serie di caratteristiche che qui tralasciamo
- In ambito digitale possiamo considerare un font come una collezione indicizzata di glifi
- Può contenere regole di metrica (kerning), composizione (segni diacritici e legature: fi), sostituzione condizionale dei glifi (script arabi), etc.
- Bitmap / Vector (o outline)
- TrueType, OpenType, Postscript Type 1, SVG, Web Open Font Format (woff), Metafont (TeX) e molti altri
- Uso improprio dei font per rappresentare caratteri non disponibili in un determinato sistema (es.: font Symbol in Windows per il greco)
- Fonts & Encodings, Yannis Haralambous, O'Really 2007 (ISBN 978-0-596-10242-5)
- http://www.microsoft.com/typography/default.mspx

Fonts & Encodings



Script

- Si può intendere:
  - Un particolare sistema di scrittura (arabo, greco, italiano, etc..)
  - L'insieme dei caratteri necessari in quel sistema di scrittura (ad es.: lo script arabo, lo script devangari, etc..)
- Un sistema di scrittura implica dei grafemi (caratteri rappresentabili), e regole di scrittura, di utilizzo e di formattazione
- A volte si parla di script per significare un'intera famiglia (lo script latino, lo script CJK, etc..)
- Tipologie di script: alfabetico LTR, alfabetico RTL, alfasillabico, sillabico, ideografico, geroglifico, cuneiforme, etc..
- Si distingue dalla lingua, nel senso che un testo scritto in una certa lingua può contenere più sistemi di scrittura contemporaneamente (in giapponese: hiragana, katakana, han, latin)
- http://www.unicode.org/charts/
- ISO 15924 standardizza i nomi degli script
- Per saperne di più:
  - http://www.unicode.org/iso15924/
  - http://www.unicode.org/reports/tr24/
  - http://www.omniglot.com/writing/

Definizioni

- Character repertoire
- Character code
- Character encoding

Character repertoire (repertorio)

- Insieme di caratteri distinti.
- Non assume alcuna rappresentazione digitale interna o finalizzata alla scambio dati.
- Il repertorio non definisce neanche un ordinamento.
- Lo si definisce tramite un elenco dei nomi dei caratteri e una rappresentazione esemplificativa degli stessi.
- Il repertorio può contenere caratteri che “sembrano” gli stessi in alcune rappresentazioni tipiche ma che sono logicamente distinti (come la A maiuscola latina, quella cirillica e la alpha maiuscola greca)

Character code (codice)

- Una mappatura, spesso presentata in forma tabulare, tra i caratteri del repertorio e un insieme di interi non negativi
- Assegna ad ogni carattere un codice numerico (code position, o code point)
- I numeri assegnati non devono necessariamente essere consecutivi
- Implica (ovviamente) la definizione del repertorio

Character encoding (codifica)

- Un metodo (algoritmo) per presentare i caratteri in forma digitale, mappando i code point in sequenze di ottetti.
- Gli encoding hanno un nome, che può essere registrato (IANA)
- Implica la definizione di un repertorio e di un character code
- http://www.unicode.org/reports/tr17/
- *character set* e *charset*: terminologie che introducono confusione. Di fatto si tratta di *encoding*.

NUL	Null char	0	00	+	Plus	43	2B	V	Uppercase V	86	56
SOH	Start of Heading	1	01	,	Comma	44	2C	W	Uppercase W	87	57
STX	Start of Text	2	02	-	Hyphen	45	2D	X	Uppercase X	88	58
ETX	End of Text	3	03	.	Period	46	2E	Y	Uppercase Y	89	59
EOT	End of Transmission	4	04	/	Slash	47	2F	Z	Uppercase Z	90	5A
ENQ	Enquiry	5	05	0	Zero	48	30	[	Opening bracket	91	5B
ACK	Acknowledgment	6	06	1	One	49	31	\	Backslash	92	5C
BEL	Bell	7	07	2	Two	50	32	]	Closing bracket	93	5D
BS	Back Space	8	08	3	Three	51	33	^	Caret	94	5E
HT	Horizontal Tab	9	09	4	Four	52	34	_	Underscore	95	5F
LF	Line Feed	10	0A	5	Five	53	35	`	Grave accent	96	60
VT	Vertical Tab	11	0B	6	Six	54	36	a	Lowercase a	97	61
FF	Form Feed	12	0C	7	Seven	55	37	b	Lowercase b	98	62
CR	Carriage Return	13	0D	8	Eight	56	38	c	Lowercase c	99	63
SO	Shift Out / X-On	14	0E	9	Nine	57	39	d	Lowercase d	100	64
SI	Shift In / X-Off	15	0F	:	Colon	58	3A	e	Lowercase e	101	65
DLE	Data Line Escape	16	10	;	Semicolon	59	3B	f	Lowercase f	102	66
DC1	Device Control 1 (XON)	17	11	<	Less than	60	3C	g	Lowercase g	103	67
DC2	Device Control 2	18	12	=	Equals	61	3D	h	Lowercase h	104	68
DC3	Device Control 3 (XOFF)	19	13	>	Greater than	62	3E	i	Lowercase i	105	69
DC4	Device Control 4	20	14	?	Question mark	63	3F	j	Lowercase j	106	6A
NAK	Negative Ack.	21	15	@	At symbol	64	40	k	Lowercase k	107	6B
SYN	Synchronous Idle	22	16	A	Uppercase A	65	41	l	Lowercase l	108	6C
ETB	End of Transmit Block	23	17	B	Uppercase B	66	42	m	Lowercase m	109	6D
CAN	Cancel	24	18	C	Uppercase C	67	43	n	Lowercase n	110	6E
EM	End of Medium	25	19	D	Uppercase D	68	44	o	Lowercase o	111	6F
SUB	Substitute	26	1A	E	Uppercase E	69	45	p	Lowercase p	112	70
ESC	Escape	27	1B	F	Uppercase F	70	46	q	Lowercase q	113	71
FS	File Separator	28	1C	G	Uppercase G	71	47	r	Lowercase r	114	72
GS	Group Separator	29	1D	H	Uppercase H	72	48	s	Lowercase s	115	73
RS	Record Separator	30	1E	I	Uppercase I	73	49	t	Lowercase t	116	74
US	Unit Separator	31	1F	J	Uppercase J	74	4A	u	Lowercase u	117	75
	Space	32	20	K	Uppercase K	75	4B	v	Lowercase v	118	76
!	Exclamation mark	33	21	L	Uppercase L	76	4C	w	Lowercase w	119	77
"	Double quotes	34	22	M	Uppercase M	77	4D	x	Lowercase x	120	78
#	Number	35	23	N	Uppercase N	78	4E	y	Lowercase y	121	79
$	Dollar	36	24	O	Uppercase O	79	4F	z	Lowercase z	122	7A
%	Percent sign	37	25	P	Uppercase P	80	50	{	Opening brace	123	7B
&	Ampersand	38	26	Q	Uppercase Q	81	51	|	Vertical bar	124	7C
'	Single quote	39	27	R	Uppercase R	82	52	}	Closing brace	125	7D
(	Open parenthesis	40	28	S	Uppercase S	83	53	~	Tilde	126	7E
)	Close parenthesis	41	29	T	Uppercase T	84	54		Delete	127	7F
*	Asterisk	42	2A	U	Uppercase U	85	55				

Esempio: Ascii

1. Si definisce il repertorio tramite nomi ed esempi
2. Si assegna un numero a ciascun carattere
3. Si definisce un algoritmo che assegna una sequenza di ottetti a ciascun numero

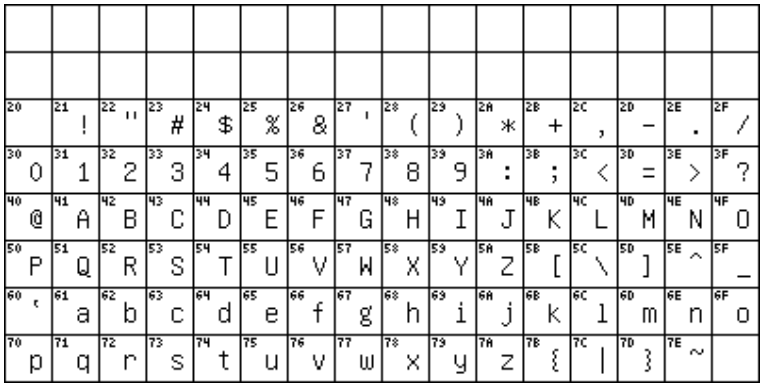
Esempi

M ! è € 中 	È un insieme di caratteri che potrebbe definire un repertorio
M !	Sono caratteri che fanno parte del repertorio definito da ASCII e da molti altri encoding
M ! è	Fanno parte del repertorio definito da ISO-8859-1 e da molti altri encoding
M ! è €	Fanno parte del repertorio definito da Windows 1252 (e altri, tra cui ISO-8859-15)
M ! è 中	Fanno parte del repertorio definito da GB 2312
M ! è € 中 	Fanno parte del repertorio definito da Unicode

		M	!	è	€	中	
ASCII	Code point (decimale)	77	33				
	Encoding (Byte)	4D	21				
Latin1	Code point (decimale)	77	33	232			
	Encoding (Byte)	4D	21	E8			
Windows 1252	Code point (decimale)	77	33	232	128		
	Encoding (Byte)	4D	21	E8	80		
GB 2312	Code point (decimale)	77	33	808		5448	
	Encoding EUC-CN (Byte)	4D	21	A8A8		D6D0	
Unicode	Code point (decimale)	77	33	232	8364	20013	77952
	Encoding UTF-16 BE (Byte)	004D	0021	00E8	20AC	4E2D	D80C DC80
	Encoding UTF-8 (Byte)	4D	21	C3A8	E282AC	E4B8AD	F0938280

Prospettiva storica

- Codice Morse (1835/1837)
- Codice Baudot (1874)
- Codice Murray (1899/1900)
- Codice EBCDIC (1963/64)
- Codice ASCII (1963/67)
- Standard ISO 8859 (anni ‘80)
- Standard Proprietari (anni ‘60 – 2000)
- Unicode (anni ‘90)
- http://www.wps.com/projects/codes/
- http://tronweb.super-nova.co.jp/characcodehist.html



20	21 !	22 "	23 #	24 $	25 %	26 &	27 '	28 (	29)	2A *	2B +	2C ,	2D -	2E .	2F /
30 0	31 1	32 2	33 3	34 4	35 5	36 6	37 7	38 8	39 9	3A :	3B ;	3C <	3D =	3E >	3F ?
40 @	41 A	42 B	43 C	44 D	45 E	46 F	47 G	48 H	49 I	4A J	4B K	4C L	4D M	4E N	4F O
50 P	51 Q	52 R	53 S	54 T	55 U	56 V	57 W	58 X	59 Y	5A Z	5B [	5C \	5D]	5E ^	5F _
60 `	61 a	62 b	63 c	64 d	65 e	66 f	67 g	68 h	69 i	6A j	6B k	6C l	6D m	6E n	6F o
70 p	71 q	72 r	73 s	74 t	75 u	76 v	77 w	78 x	79 y	7A z	7B {	7C |	7D }	7E ~	

Il codice ASCII

- American Standard Code for Information Interchange
- Denota un repertorio, un codice e una codifica
- La maggior parte dei character code usati oggi includono ASCII come sottoinsieme
- A causa del suo largo utilizzo in passato spesso ASCII è usato come sinonimo di “formato testo”
- La definizione di ASCII specifica dei caratteri di controllo (00-19, 7F) il cui uso è vario nonostante la standardizzazione dei nomi e delle funzioni
- La codifica mappa ogni code point nell’ottetto che ha lo stesso valore numerico
- 7 bit: gli ottetti da 128 a 255 non sono usati in ASCII

Varianti nazionali di ASCII (ISO 646)

- Esistono varianti *nazionali* di ASCII in cui alcuni caratteri speciali sono stati rimpiazzati da caratteri più comuni in un'altra lingua
- La formulazione *originale* di ASCII viene perciò denominata spesso US-ASCII. N.B.: parlare di formulazione originale è improprio dal momento che lo standard ha subito varie modifiche tra il 1963 e il 1968 (ANSI_X3.4-1968)
- **ISO 646** definisce un set di caratteri simile ad ASCII in cui le posizioni occupate in ASCII dai caratteri @[\]{|} sono assegnate per uso nazionale.
- #$^`~ possono anche essere usati se necessario
- Quasi tutti i caratteri utilizzati nelle varianti nazionali sono stati inclusi in ISO-8859-1 (Latin1)

Varianti nazionali di ASCII

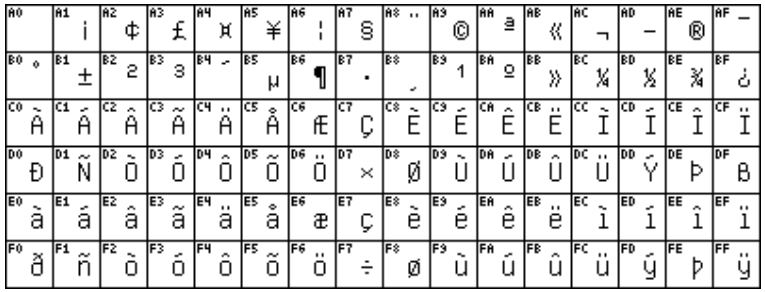
dec	hex	glifo	Nome ufficiale Unicode	National variants
35	23	#	NUMBER SIGN	£ Ù
36	24	$	DOLLAR SIGN	¤
64	40	@	COMMERCIAL AT	É § Ä à ³
91	5B	[	LEFT SQUARE BRACKET	Ä Æ ° â ¡ ÿ é
92	5C	\	REVERSE SOLIDUS	Ö Ø ç Ñ ½ ¥
93	5D	]	RIGHT SQUARE BRACKET	Å Ü § ê é ¿ |
94	5E	^	CIRCUMFLEX ACCENT	Ü î è
96	60	`	GRAVE ACCENT	é ä µ ô ù
123	7B	{	LEFT CURLY BRACKET	ä æ é à ° ¨
124	7C	|	VERTICAL LINE	ö ø ù ò ñ *f*
125	7D	}	RIGHT CURLY BRACKET	å ü è ç ¼
126	7E	~	TILDE	ü ¯ ß ¨ û ì ´ _

Caratteri ASCII “sicuri”

- A causa dell’esistenza delle varianti nazionali di ASCII alcuni caratteri sono meno *sicuri* di altri
- Oltre alle lettere dell’alfabeto inglese (da “A” a “Z”, da “a” a “z”), le cifre (da “0” a “9”) e lo spazio, i caratteri che possono essere considerati sicuri nella trasmissione dei dati sono i seguenti:

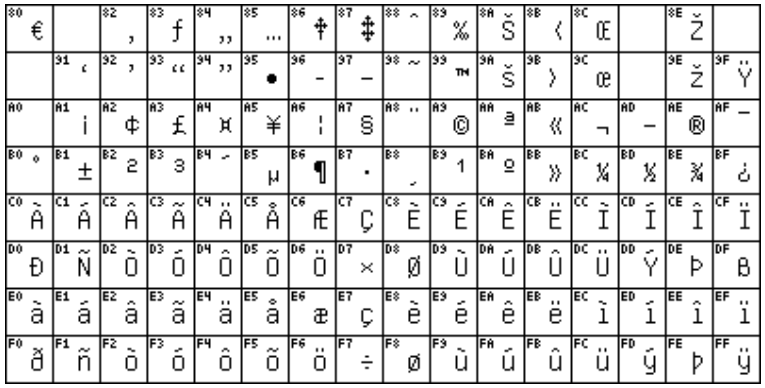
```
! " % & ' ( ) * + , - . / : ; < = > ? _
```

- Si noti che alcuni di questi caratteri possono essere interpretati in maniera particolare dal destinatario (sia umano che software)
- Esistono comunque encoding che mappano questi caratteri in altri ottetti (ad es.: EBCDIC, GMS)



ISO Latin 1 alias ISO 8859-1

- ISO-8859-1, parte della famiglia ISO-8859, definisce un repertorio di caratteri denominato “Latin alphabet No. 1”
- Lo standard definisce anche un codice e un encoding in maniera simile ad ASCII: ogni code point è mappato semplicemente ad un ottetto che ha lo stesso valore numerico del code point (tutti e 8 i bit)
- Oltre ai caratteri ASCII, ISO Latin 1 contiene vari caratteri composti con segni diacritici, necessari alla scrittura delle lingue dell’Europa occidentale, e ulteriori caratteri speciali
- Il primo carattere nella figura è il *no-break space*
- Posizioni da 80 a 9F riservate (C1 controls)
- http://www.cs.tut.fi/~jkorpela/latin1/index.html



Windows Latin 1 (cp-1252)

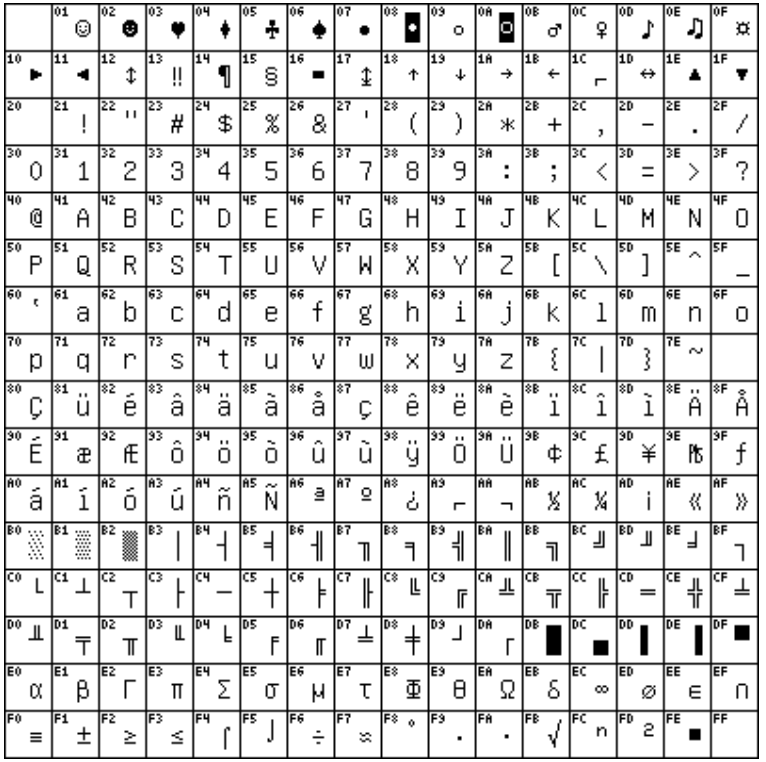
- Codifica proprietaria Microsoft
- Sfrutta le posizioni riservate da ISO-8859-1 (80-9F) per inserire ulteriori caratteri stampabili
- Spesso chiamata “ANSI” (di fatto è un errore, ma molti programmi usano questa terminologia, compreso Notepad++)
- Negli stessi sistemi Windows alcuni programmi potrebbero usare altre codifiche (DOS code pages)
- Estende di fatto Latin1 con molti caratteri tipograficamente importanti (smart quotes, em ed en dash) che spesso troviamo con i copia e incolla da Word nei Wysiwig dei CMS
- La processazione di un testo in Windows Latin 1 da parte di un programma che si aspetta come input del testo in ISO-8859-1 può produrre risultati inaspettati (sempre meno)
- Esistono diversi Windows charset (o code pages, CP) che differiscono dai corrispondenti ISO-8859-1 nelle posizioni riservate (80-9F) (con alcuni distinguo)

La famiglia ISO 8859

- Analogamente all'estensione di ASCII da parte di ISO-8859-1 e Windows Latin1 sono state standardizzate molte altre estensioni a 8 bit, la più importante delle quali è ISO 8859
- Attualmente esistono 15 parti dell'ISO 8859
- Ad esempio: ISO-8859-2 estende ASCII con i caratteri necessari nelle lingue dell'Europa centrale e orientale
- 80-9F sono posizioni riservate in tutto ISO 8859
- Si utilizza sempre lo stesso encoding (ottetto con lo stesso valore numerico)
- ISO-8859-15 alias Latin9 (introduzione del simbolo dell'Euro e correzioni a Latin1) è stato un "fallimento"
- http://www.cs.tut.fi/~jkorpela/8859.html

Altre estensioni ad ASCII

- DOS character codes o code pages
  - CP 437: codepage originale americano che conteneva alcune lettere greche, simboli matematici, simboli per il disegno di tabelle in formato testo e altre amenità (smilies)
  - CP 850: come ISO-8859-1, conteneva i caratteri necessari per le lingue occidentali (in altre posizioni)
  - La Microsoft oggi li chiama *OEM code pages* (per aumentare la confusione)
- Macintosh character codes
- HP, Adobe, CJK (codifiche asiatiche), medio oriente, archeo, etc…



http://www.iana.org/assignments/character-sets

GSM 03.38

Il character encoding standard per i messaggi GSM è una codifica a 7 bit.
L'encoding non è "banale": per alcuni caratteri occorre utilizzare un "settetto" speciale di prefisso (ESC).
Non è una estensione ad ASCII, anche se i caratteri A-Z,a-z,0-9 (e qualche altro) hanno gli stessi code point (e sono codificati dagli stessi 7 bit).

Basic Character Set

	0x00	0x10	0x20	0x30	0x40	0x50	0x60	0x70
0x00	@	Δ	SP	0	¡	P	¿	p
0x01	£	_	!	1	A	Q	a	q
0x02	$	Φ	"	2	B	R	b	r
0x03	¥	Γ	#	3	C	S	c	s
0x04	è	Λ	¤	4	D	T	d	t
0x05	é	Ω	%	5	E	U	e	u
0x06	ù	Π	&	6	F	V	f	v
0x07	ì	Ψ	'	7	G	W	g	w
0x08	ò	Σ	(	8	H	X	h	x
0x09	Ç	Θ	)	9	I	Y	i	y
0x0A	LF	Ξ	*	:	J	Z	j	z
0x0B	Ø	ESC	+	;	K	Ä	k	ä
0x0C	ø	Æ	,	<	L	Ö	l	ö
0x0D	CR	æ	-	=	M	Ñ	m	ñ
0x0E	Å	ß	.	>	N	Ü	n	ü
0x0F	å	É	/	?	O	§	o	à

- LF is a Line Feed control.
- CR is a Carriage Return control, or filler.
- ESC is an Escape control.
- SP is a Space character.

Basic Character Set Extension

	0x00	0x10	0x20	0x30	0x40	0x50	0x60	0x70
0x00					|			
0x01								
0x02								
0x03								
0x04		^						
0x05							€	
0x06								
0x07								
0x08			{					
0x09			}					
0x0A	FF							
0x0B		SS2						
0x0C				[				
0x0D	CR2			~				
0x0E				]				
0x0F			\					

- FF is a Page Break control. If not recognized, it shall be treated like LF.
- CR2 is a control character. No language specific character shall be encoded at this position.
- SS2 is a second Single Shift Escape control reserved for future extensions.

Negli SMS oggi esiste(rebbe) anche la possibilità di utilizzare:

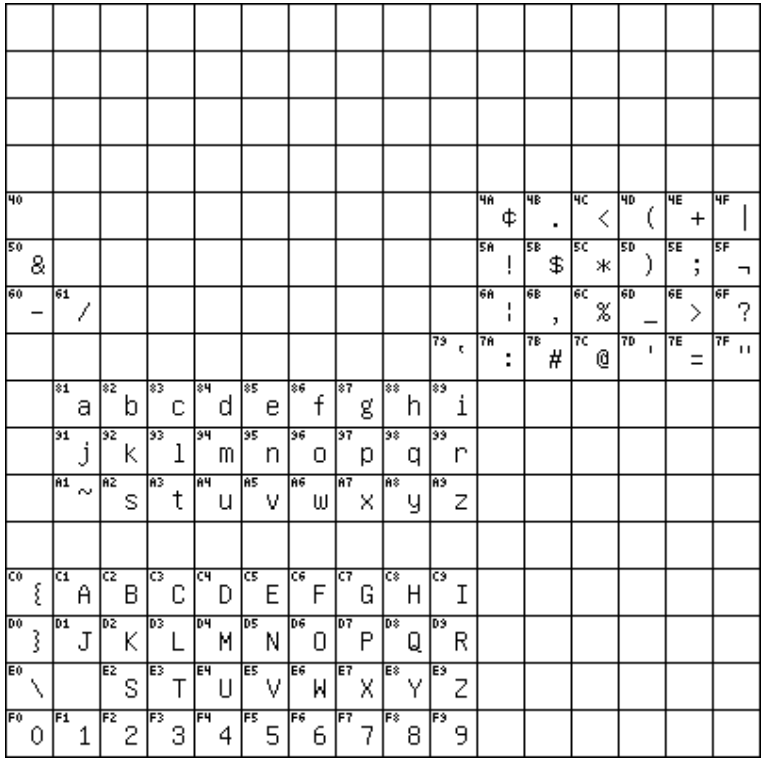
- GSM a 8 bit (140 byte)
- codifica UCS-2 (70 caratteri)
- National language shift tables (*spagnolo, portoghese, turco, 10 lingue indiane con script Brahmi, inglese esteso, tedesco, olandese, svedese, danese, finlandese, norvegese, francese, italiano, ungherese, polacco, ceco, islandese, greco, russo, ebraico e arabo*)

Altri codici a 8 bit

- Tutte le codifiche fin qui presentate (a parte GSM 03.38) sono codifiche a 8 bit in cui l'algoritmo di codifica è banale (ottetto = code point)
- Ci sono sempre 256 code points, alcuni dei quali sono riservati come codici di controllo o lasciati inutilizzati
- Sebbene la maggior parte delle codifiche a 8 bit sia una estensione ad ASCII, ciò è dovuto principalmente al largo uso di ASCII precedente alla definizione della nuova codifica
- Gli standard ISO 2022 e ISO 4873 definiscono un framework generale per codici a 7 e 8 bit e per "switchare" fra i codici. L'idea è di utilizzare le posizioni C1 controls (80-9F) a cui però i codici Windows non si attengono
- EBCDIC è una codifica a 8 bit definita dalla IBM nei primi anni '60 e tutt'ora utilizzata in alcuni mainframe.
- EBCDIC contiene TUTTI i caratteri ASCII, ma in posizioni diverse
- Una dettaglio degno di nota è che in EBCDIC le lettere A-Z non appaiono in posizioni consecutive
- Anche EBCDIC esiste in diverse varianti nazionali
- http://www.terena.org/activities/multiling/euroml/section05.html

IBM EBCDIC

- Extended Binary Coded Decimal Interchange Code (EBCDIC)



ISO 10646, UCS e Unicode

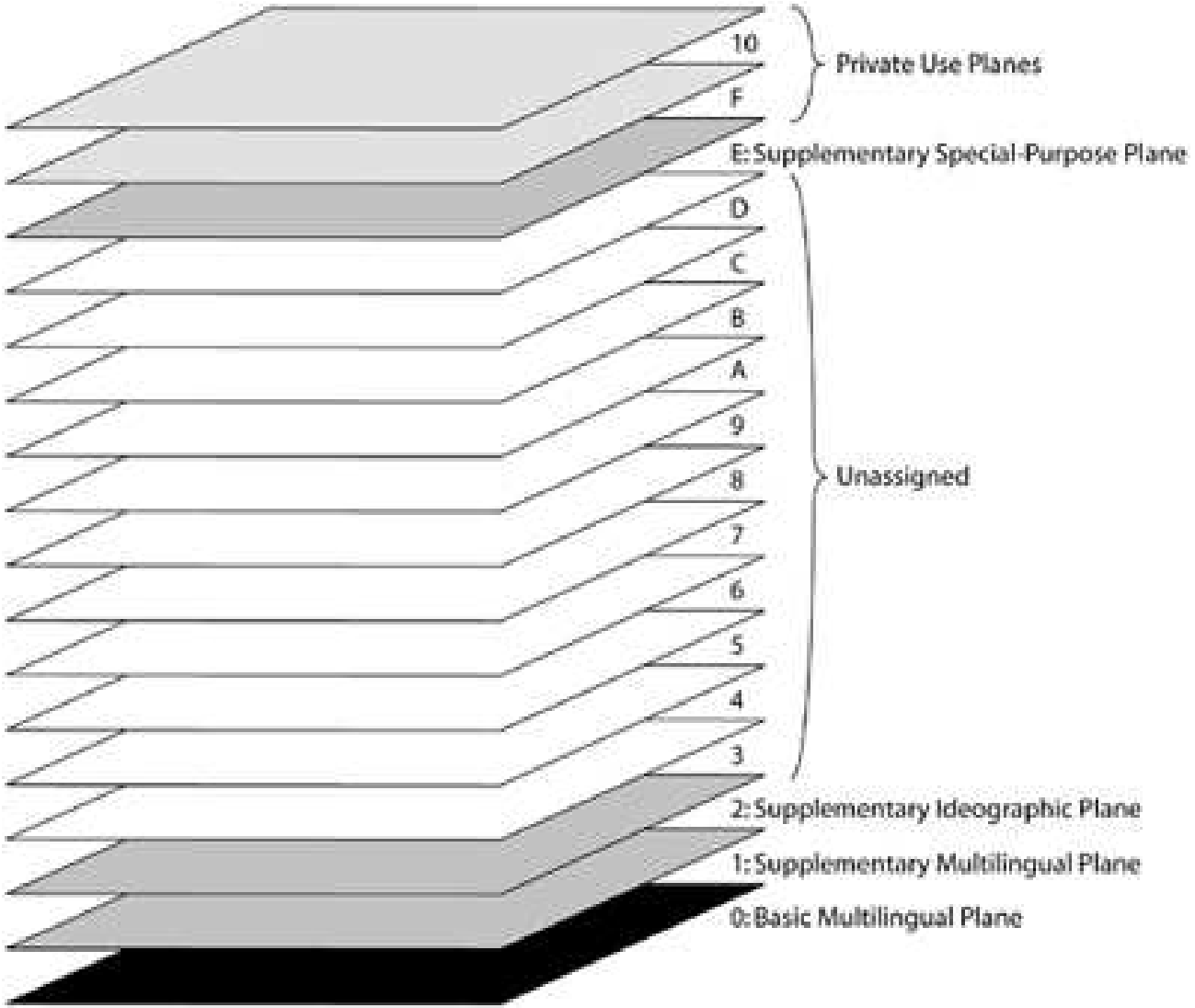
- ISO 10646 è uno standard ufficiale internazionale
  - Originariamente prevede 32 bit per lo spazio di indirizzamento
  - Definisce l'UCS (Universal Character Set) e due codifiche
  - È in crescita costante e contiene tutti i caratteri definiti in tutte le altre codifiche utilizzate

- Unicode
  - Originariamente prevede 16 bit per lo spazio di indirizzamento
  - È lo standard definito dall'Unicode Consortium
  - Il repertorio e il codice sono completamente compatibili con ISO 10646 (l'accordo è di usare solo 17 "piani" a 16 bit)
  - Unicode aggiunge molti aspetti tecnico-pratici
  - Unicode definisce più codifiche (encoding)

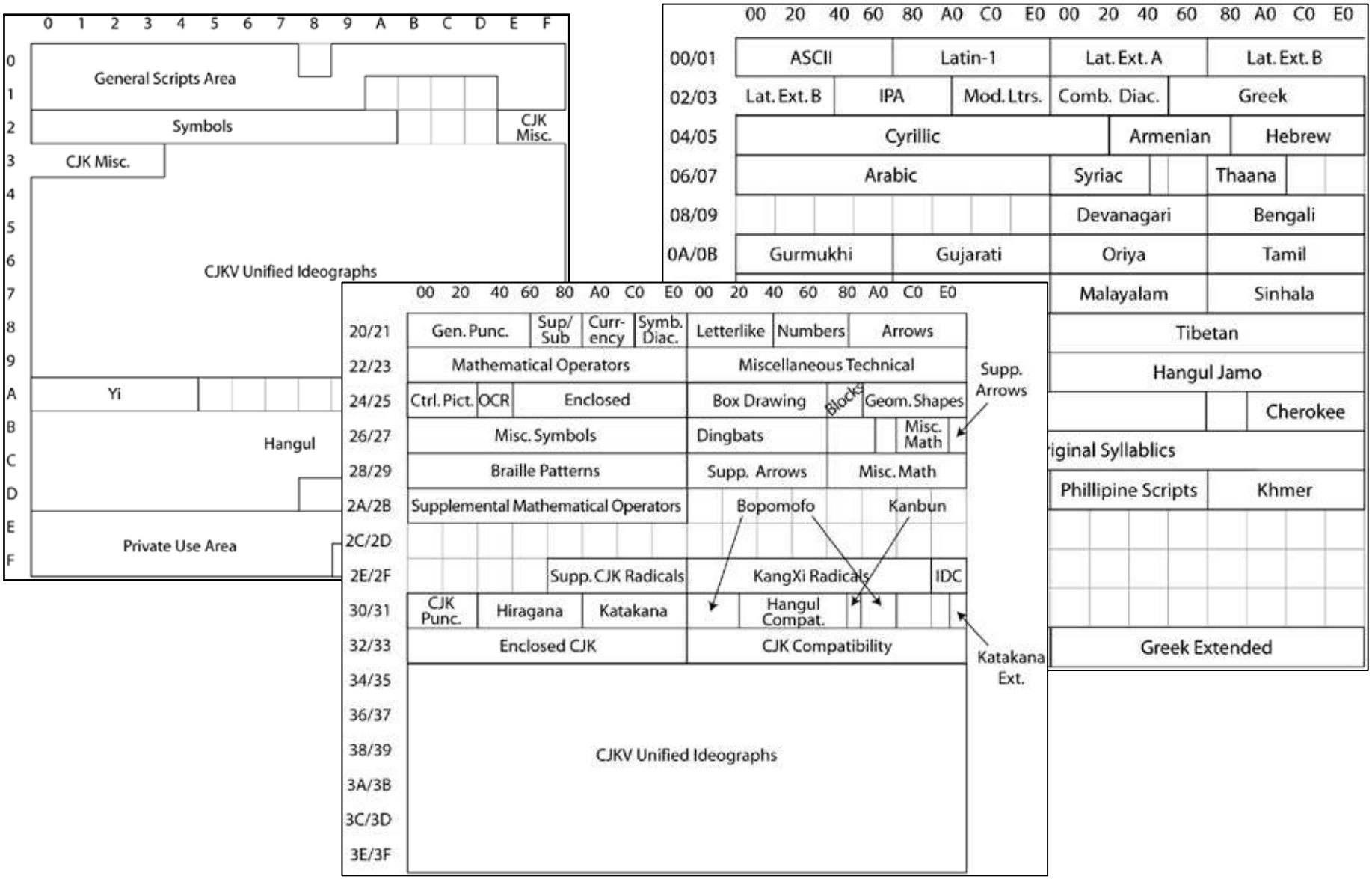
Unicode

- Si parte dai code points piuttosto che dai caratteri
- Originariamente (Unicode 1.0, 1992) era un codice a 16 bit: 65.536 code points
- È stato esteso e suddiviso in 17 “piani” numerati da 0 a 16 in cui ciascun piano è uno spazio di indirizzamento di 16 bit
- Per 17 “piani” occorrono almeno 5 bit supplementari: 16 + 5 = 21 bit
- I 21 bit non sono utilizzati tutti: 65.536 × 17 = 1.114.112 ($< 2^{21}$ = 2.097.152)
- I code points sono interi nel range esadecimale 0.. 10FFFF (0.. 1.114.111)
- Fino a tempi molto recenti l’uso di Unicode si è limitato al BMP (Basic Multilingual Plane) nel range 0.. FFFF (i primi 16 bit)
- Attualmente (Unicode 6.2) sono definiti 110.117 caratteri rappresentabili
  http://www.babelstone.co.uk/Unicode/unicode.html

Piani Unicode



Piano BMP


0 1 2 3 4 5 6 7 8 9 A B C D E F
General Scripts Area
Symbols
CJK Misc.
CJK Misc.
CJKV Unified Ideographs
Yi
Hangul
Private Use Area
00 20 40 60 80 A0 C0 E0 00 20 40 60 80 A0 C0 E0
00/01 ASCII Latin-1 Lat. Ext. A Lat. Ext. B
02/03 Lat. Ext. B IPA Mod. Ltrs. Comb. Diac. Greek
04/05 Cyrillic Armenian Hebrew
06/07 Arabic Syriac Thaana
08/09 Devanagari Bengali
0A/0B Gurmukhi Gujarati Oriya Tamil
Malayalam Sinhala
Tibetan
Hangul Jamo
Cherokee
iginal Syllablics
Phillipine Scripts Khmer
Greek Extended
00 20 40 60 80 A0 C0 E0 00 20 40 60 80 A0 C0 E0
20/21 Gen. Punc. Sup/ Sub Curr- ency Symb. Diac. Letterlike Numbers Arrows
22/23 Mathematical Operators Miscellaneous Technical
24/25 Ctrl. Pict. OCR Enclosed Box Drawing Blocks Geom. Shapes
26/27 Misc. Symbols Dingbats Misc. Math
28/29 Braille Patterns Supp. Arrows Misc. Math
2A/2B Supplemental Mathematical Operators
2C/2D
2E/2F Supp. CJK Radicals KangXi Radicals IDC
30/31 CJK Punc. Hiragana Katakana Hangul Compat.
32/33 Enclosed CJK CJK Compatibility
34/35
36/37
38/39 CJKV Unified Ideographs
3A/3B
3C/3D
3E/3F
Supp. Arrows
Bopomofo
Kanbun
Katakana Ext.

Blocchi

Name	From	To	# Codepoints
Basic Latin	U+0000	U+007F	(128)
Latin-1 Supplement	U+0080	U+00FF	(128)
Latin Extended-A	U+0100	U+017F	(128)
Latin Extended-B	U+0180	U+024F	(208)
IPA Extensions	U+0250	U+02AF	(96)
Spacing Modifier Letters	U+02B0	U+02FF	(80)
…	…	…	…
…	…	…	…
Mongolian	U+1800	U+18AF	(156)
…	…	…	…
CJK Unified Ideographs	U+4E00	U+9FFF	(20941)
…	…	…	…
Supplementary Private Use Area-B	U+100000	U+10FFFF	(2)

http://www.fileformat.info/info/unicode/block/index.htm

http://www.unicode.org/Public/UNIDATA/Blocks.txt

http://www.babelstone.co.uk/Unicode/babelmap.html